STATISTICS IN MEDICINE, VOL. 13, 651–661 (1994)

EPIDEMIOLOGIC METHODS

A COMPARISON OF THREE APPROACHES TO ESTIMATE EXPOSURE-SPECIFIC INCIDENCE RATES FROM POPULATION-BASED CASE-CONTROL DATA

JACQUES BENICHOU AND SHOLOM WACHOLDER

National Cancer Institute, Biostatistics Branch, 6130 Executive Blvd., EPN/403, Rockville, Maryland 20892, U.S.A.

SUMMARY

In population-based case-control studies, an attempt is made to identify all incident cases diagnosed in a specified population during a fixed time interval. Assuming that this goal is met allows one to obtain measures of risk other than relative risks. In this paper, we describe three approaches to estimate exposure-specific incidence rates. Approach 1 relies on estimating crude incidence rates of the disease in strata defined, for instance, by age and geographic area, and combining them with relative risk estimates from the case-control data. In approaches 2 and 3, baseline incidence rates and relative risks are estimated jointly. Approach 2 is based on a pseudo-likelihood, while, in approach 3, the problem is regarded as a missing data problem and a full likelihood is maximized. We applied these three approaches to a study of bladder cancer. Our three sets of estimates of exposure-specific incidence rates were in close agreement, while there appeared to be greater precision with approaches 2 and 3.

1. INTRODUCTION

Case-control studies provide data on the distributions of exposure respectively in diseased subjects (cases) and non-diseased subjects (controls) for the disease under study. These data are used to estimate relative risks through the estimation of odds ratios. In population-based case-control studies, an effort is made to identify all incident cases diagnosed in a specified population during a fixed time interval. In principle, then, other measures of risk can be estimated, exploiting the counts of incident cases and of persons or person-years at risk for the disease in strata or subsets of the population, defined perhaps by age, race, sex and geographic area.

To be specific, crude incidence rates of the disease can be calculated in each stratum as the ratio of the number of incident cases to the total number at risk. Further, by combining the crude incidence rate and relative risk estimates, one can obtain, in each stratum, exposure-specific estimates of disease incidence rates, that is incidence rates for all exposure combinations, as has been noted by Cornfield,[1,2] MacMahon,[3] Miettinen,[4] Neutra and Drolette,[5] Greenland[6] and Gail *et al.*[7] However, methods of inference for the exposure-specific incidence rates have not been fully developed, particularly in the multivariate setting.

In this paper, we present three approaches for estimating exposure-specific incidence rates from population-based case-control studies. The results of application of these three approaches to a population-based case-control study of bladder cancer[8] are compared. In Section 2, we describe

CCC0277–6715/94/070651–11

this study. We then describe the three approaches in Section 3 and apply them to the bladder cancer data in Section 4. Section 5 provides a discussion.

2. THE NATIONAL BLADDER CANCER STUDY

The National Bladder Cancer Study was a population-based case-control study conducted at ten cancer registries in the United States, nine of which were part of the Surveillance, Epidemiology and End Results Program of the National Cancer Institute. All individuals aged 21–84 years with a first histologically confirmed bladder cancer diagnosed during the year 1978 and reported to the tumour registries were asked to be cases. Controls aged 21–84 years were frequency-matched to cases on geographic area, age (in five-year intervals) and sex. Approximately twice as many controls as cases were selected except in New Jersey where the ratio was 3:2 and in Detroit where it was only 1:1. Controls aged 21–64 years were selected from 22,633 households chosen by telephone sampling with the use of random-digit dialing while controls aged 65–84 years were selected from Health Care Financing Administration rosters.[8]

A total of 4086 eligible cases and 6985 controls were identified. Among them, respectively 2982 (73 per cent) cases and 5782 (83 per cent) controls were interviewed. Information on exposure to several potential risk factors was gathered. They included cigarette smoking, occupational exposure (to a list of substances including dye, rubber, leather, ink or paint), history of urinary tract infection, coffee drinking, tap water consumption and artificial sweeteners consumption. For the purpose of this illustrative analysis, we will consider only the two main risk factors, namely cigarette smoking, classified into four levels: never smoked; ex-smoker; current light smoker (< 20 cigarettes/day); and current heavy smoker ($\geqslant 20$ cigarettes/day); and occupational exposure into two levels. An occupation was considered to be at high risk if any of the following criteria was met:

1. on a list of suspected occupations based on the previous literature and with an observed relative risk in this study of $\geqslant 1{\cdot}3$;
2. any occupation with a relative risk in this study of $\geqslant 1{\cdot}5$; or
3. an occupation demonstrating a (significant) positive trend in risk in this study with increasing number of years worked.[9]

Moreover, we only consider three out of the ten geographic areas, with low, moderate and high crude incidence rates, respectively, the areas covered by the registries for Utah, Atlanta, and New Jersey. We will consider nine strata defined by geographic area and age group (21–44 years, 45–64 years, 65+ years). Only white men, the largest race-sex group, are included in our analyses.

Because the study was population-based, the cases could be linked to the population from which they came. The distribution of age, sex and race in each area in the year 1978 was extrapolated from U.S. Census data for the years 1970 and 1980, and provided by the SEER Program.[10] These counts provide a close approximation to the number of person-years in each age-area stratum since the cases were diagnosed during a one-year interval and the population could be considered to be stable during that time period. Therefore, the crude incidence rates could be calculated from a full census of incident cases and the number of person-years in each stratum defined by age and geographic area. Table I presents the number of cases and controls for whom we had information on exposure as well as the total number of incident cases and subjects at risk during 1978 for each stratum. Note that in the rest of the paper, we will use the term cases (and controls) for diseased (and non-diseased) persons with information on exposure.

Table I. Bladder cancer data (white males)

	Population in 1978	Number of 1978 incident cases	Number of cases with information on exposure	Number of controls with information on exposure
Utah				
age (years)				
21–44	238,184	9	9	24
45–64	97,420	32	23	83
65+	40,236	70	48	91
Atlanta				
age (years)				
21–44	247,185	9	7	8
45–64	110,095	53	33	66
65+	31,662	71	28	68
New Jersey				
age (years)				
21–44	300,515	39	28	62
45–64	216,789	309	246	510
65+	91,579	531	336	607
Total	1,373,665	1123	758	1519

3. THE THREE APPROACHES FOR ESTIMATING EXPOSURE-SPECIFIC INCIDENCE RATES

We now describe three approaches for estimating exposure-specific incidence rates in population-based case-control studies such as the National Bladder Cancer Study. These approaches are presented in a general fashion but specifics for the bladder cancer data are pointed out when necessary. First, we describe the basic models involved for the three approaches.

3.1. Models

Let X denote the exposure factors influencing the probability of occurrence of the disease under study and S the stratification factors. We restrict ourselves to X and S being categorical variables, respectively indexed by i ($i = 1, 2, \ldots, I$) and j ($j = 1, 2, \ldots, J$). Summation over an index (excluding value 0) is indicated by dot notation. In our analysis of the bladder cancer data, X represents cigarette smoking and occupational exposure, and S represents geographic area and age. Therefore, there are eight levels of combined cigarette smoking and occupation exposure ($I = 8$) and nine strata defined by geographic area and age ($J = 9$).

The target population of the case-control study is considered to be followed for a set time period for the occurrence of bladder cancer. As in cohort studies that model incidence rates with complete data on exposure,[11] we can assume that the number of observed incident cases d_{ij} in joint level $i \times j$ of $X \times S$ has a Poisson distribution with mean $h_{ij}t_{ij}$, where t_{ij} and h_{ij} are, respectively, the follow-up time on study and the disease rate in joint level $i \times j$ of $X \times S$. In our example, the t_{ij} are unknown quantities, since the number of individuals at each level of X in the jth stratum is unknown, but each $t_{.j}$, the follow-up time for the jth stratum, is assumed known, and set equal to the census counts of individuals in the stratum. We used 1978 counts since 1978 was the year of diagnosis for all the cases. Inevitably, the value of X will not be known for all identified cases (due to refusal to participate or death), but typically S will be known even for these cases.

Instead of using the Poisson model, we can assume that the d_{ij} are binomial with mean parameters p_{ij} and total individuals t_{ij}. Then, the p_{ij} represent the risk of disease during follow-up. For our bladder cancer example, where the disease rate is low for all joint levels of i and j and follow-up is for at most one year, the Poisson and binomial models are approximately equivalent and the choice between them is a matter of convenience. The three approaches could be implemented only with the binomial assumption or, alternatively, only with the Poisson assumption (except for the case-control part of approach 1, see below).

For approaches 1 and 2, we assume that the disease rate h_{ij} is given by the multiplicative Poisson model:

$$h_{ij} = h_{1j} R_{ij},$$

where h_{1j} is the baseline disease rate for level j of S and R_{ij} is the relative risk for joint level $i \times j$ of $X \times S$ compared to joint level $1 \times j$ of $X \times S$. Note that since R is indexed by both i and j, the proportional hazard assumption need not be made, though it is made below in our analyses of the bladder cancer data. Also note that the Poisson assumption yields the same likelihood as is obtained by assuming that time to disease has piecewise constant hazard within all levels of X and S.[12,13]

For approach 3 and the case-control part of approach 1, we assume that the probability of disease p_{ij} in joint level $i \times j$ of $X \times S$ follows a logistic model,

$$\operatorname{logit} p_{ij} = \alpha_j + X_i \beta, \tag{1}$$

where X_i is the vector of exposure variables for level i and β is the corresponding vector of parameters. Additional terms for exposure by stratum interactions could be added to the model, but this was not done in our analysis of the bladder cancer data.

For all three approaches we must assume that we have full case identification and that X is missing at random (MAR)[14] in order to estimate exposure-specific incidence rates. MAR requires that the distribution of X be the same for controls and other non-diseased individuals in the population, and the same for diseased individuals with known and unknown X, conditional on variables available for everyone, such as disease status, age and geographic area. Thus, the fact that interviewing response rates vary by age, geographic area and disease status[8] does not imply violation of MAR assumption.

3.2. Approach 1

This approach relies on earlier work by Miettinen[4] and multivariate extensions by Gail *et al.*[7] for a case-control study based on a cohort. It involves estimating crude incidence rates for each stratum and then obtaining baseline incidence rates (that is, incidence rates at the baseline level of exposure) by combining crude incidence rates estimates and attributable risk estimates. It was developed in the context of a cohort followed-up for a given time period in order to diagnose incident cases of the disease under study. Our argument above suggests that the approach of Gail *et al.*[7] can be readily applied here. It is outlined below.

Given that information on X is only available for the case-control data, baseline incidence rates h_{1j} cannot be estimated directly. However, the crude incidence rates h_j^* for a given level j of S (stratum j) can be estimated by $d_{.j}/t_{.j}$, where $d_{.j}$ and $t_{.j}$ are, respectively, the number of incident cases and follow-up time in level j of S. These two quantities can usually be obtained for a population-based case-control study. In our example, $t_{.j}$ is the population in level j of S in 1978. The crude and baseline rates are related through:[4,7]

$$h_{1j} = h_j^*(1 - AR_j),$$

where AR_j is the attributable risk in stratum j. Using results from Bruzzi *et al.*[15] on attributable risk, Gail *et al.*[7] obtained:

$$h_{1j} = h_j^* \sum_{i=1,I} \rho_{ij} R_{ij}^{-1},$$

where the summation is taken over all levels of X and ρ_{ij} is the proportion of cases at level j of S who are at level i of X. The quantities ρ_{ij} can be estimated from the cases. The relative risks R_{ij} can be estimated by odds ratios from the case-control data. Upon fitting a logistic model with a $p \times 1$ vector of parameters θ, one obtains the baseline incidence estimates:

$$\hat{h}_{1j} = (d_{.j}/t_{.j}) \sum_{i=1,I} (n_{ij}/n_{.j}) R_{ij}^{-1}(\hat{\theta}),$$

where n_{ij} is the number of cases with joint level $i \times j$ of $X \times S$, $n_{.j}$ is the total number of cases in level j of S, and $\hat{\theta}$ is the MLE of θ. If all incident cases in the population are used as cases for the case-control study, then $n_{ij} = d_{ij}, n_{.j} = d_{.j}$, and the previous estimate simplifies to:

$$\hat{h}_{1j} = (1/t_{.j}) \sum_{i=1,I} n_{ij} R_{ij}^{-1}(\hat{\theta}).$$

Finally, the estimate of incidence rates for level i of exposure ($i = 2, \ldots, I$) and stratum j are given by:

$$\hat{h}_{ij} = \hat{h}_{1j} R_{ij}(\hat{\theta}).$$

We have derived variance estimators elsewhere.[16] We present an outline here. All incidence rate estimates $\hat{h}_{ij}$ are functions of a finite number of parameters, namely the baseline incidence rates $\hat{h}_{1j}$, and the parameter estimates $\hat{\theta}$ from logistic regression. In order to estimate the variance of $\hat{h}_{ij}$, one needs to estimate the covariance matrix Σ of parameter estimates $\hat{h}_{1j}$ and $\hat{\theta}$, and then apply the delta-method.[17]

The covariance matrix Σ involves the following submatrices: the $J \times J$ covariance matrix $\Sigma_{\hat{h}_1,\hat{h}_1}$of baseline disease incidence rate estimates, the $J \times p$ covariance matrix $\Sigma_{\hat{h}_1,\hat{\theta}}$ of baseline disease incidence rate estimates with logistic parameter estimates, and the $p \times p$ covariance matrix $\Sigma_{\hat{\theta},\hat{\theta}}$ of parameter estimates $\hat{\theta}$. The matrix $\Sigma_{\hat{\theta},\hat{\theta}}$ can be estimated from the observed information matrix from the logistic model. However, the other two matrices are complex to estimate because the baseline incidence estimates are obtained by combining data from two sources, namely the case-control data and the population data. Details were worked out by Benichou and Gail.[16] They involve noting that the logistic score equations define implicit relationships between parameter estimates $\hat{\theta}$ and quantities n_{ij} defined above, and then applying a delta-method for implicitly defined random variables.[18]

3.3. Approach 2

The second approach relies on a pseudo-likelihood approach. If information on exposure were available for everybody in the population rather than just for the cases and controls, one could write a full likelihood for the entire population. With the Poisson assumption, defined above for disease rates h_{ij}, the log-likelihood would be:

$$l = \sum_{i=1,I} \sum_{j=1,J} [d_{ij} \log \{h_{ij}(\xi)\} - h_{ij}(\xi) t_{ij}].$$

where quantities d_{ij} and t_{ij} are as defined previously and the $q \times 1$ vector ξ consists of baseline incidence rates and relative risk parameters.

However, only the quantities $d_{.j}$ and $t_{.j}$ are observed, not the individual t_{ij} and d_{ij}. This approach relies on estimating these latter quantities from the observed distributions of X in the cases and controls, respectively. If one denotes disease status by D (0 for non-diseased and 1 for diseased), the estimated quantities $\tilde{d}_{ij} = d_{.j}\hat{P}(X = x_i|S = s_j, D = 1)$ and $\tilde{t}_{ij} = t_{.j}\hat{P}(X = x_i|S = s_j, D = 0)$ can be substituted into the Poisson likelihood to obtain a pseudo-likelihood whose logarithm is:

$$\tilde{l} = \sum_{i=1,I}\sum_{j=1,J} [\tilde{d}_{ij}\log\{h_{ij}(\xi)\} - \tilde{t}_{ij}h_{ij}(\xi)].$$

Maximum pseudo-likelihood estimates of ξ, that is of the baseline incidence rates and relative risk parameters, can thus be obtained. Using these pseudo-likelihood estimates, one can obtain estimates of exposure-specific incidence rates h_{ij}. Note that, strictly speaking, one should use $\hat{P}(X = x_i|S = s_j)$ instead of $\hat{P}(X = x_i|S = s_j, D = 0)$ to obtain $\tilde{t}_{ij}$, but, for an event as rare as bladder cancer in one year, the two are equivalent.

In this approach, baseline incidence rates and relative risks are jointly estimated using all the data. In approach 1, on the other hand, relative risks are estimated from the case-control data, and baseline incidence rates are estimated by combining crude incidence rates estimates obtained from the population data and attributable risks estimates obtained separately from the case-control data. This points to a potential gain in efficiency with approach 2.

The additional component of variability in the estimates of ξ due to using estimates for d_{ij} and t_{ij} can be taken into account through a 'sandwich' variance estimator. For this problem, the estimating equations are defined by the pseudo-likelihood score equations:

$$\partial\tilde{l}/\partial\xi_l = \sum_{i=1,I}\sum_{j=1,J} [\tilde{d}_{ij}\partial\log h_{ij}(\xi)/\partial\xi_l - \tilde{t}_{ij}\partial h_{ij}(\xi)/\partial\xi_l],$$

for $l = 1, \dots, q$. These q score equations are unbiased estimating equations provided $\tilde{d}_{ij}$ and $\tilde{t}_{ij}$ are unbiased estimates of quantities d_{ij} and t_{ij}, that is, provided that X is missing at random. If one denotes each of these q equations by $E(\xi_l)$, the vector of all q estimating equations by E, its $q \times q$ variance matrix by var (E), and the $q \times q$ matrix of derivatives of score equations with respect to the $q \times 1$ vector ξ by $\partial E/\partial\xi$, then the sandwich variance estimator for $\tilde{\xi}$, the pseudo-likelihood estimate of ξ, is given by Liang and Zeger:[19]

$$\text{var}(\tilde{\xi}) = (\partial E/\partial\xi)^{-1}\,\text{var}(E)(\partial E/\partial\xi)^{-1}.$$

In E, only the quantities $\tilde{d}_{ij}$ and $\tilde{t}_{ij}$ are random. The $q \times q$ matrix var(E) can be estimated by using the delta-method and considering the independent multinomial distributions of X for each level j of S among the controls and from the independent Poisson distributions of d_{ij} for each level i of X and j of S.

3.4. Approach 3

Approach 3 differs from approach 2 in the way quantities t_{ij} and d_{ij} are estimated. In approach 2, only the observed distribution of X among the controls and cases are used to estimated quantities t_{ij} and d_{ij}, respectively. In approach 3, the problem is regarded as a missing data problem and the estimates of t_{ij} and d_{ij} depend on the estimates of relative risk. The EM algorithm[14,20] can be used to estimate MLE's of ξ.

Alternatively, Newton–Raphson methods can be used to obtain a maximum likelihood estimate of the parameters of interest. Newton–Raphson converges much more quickly but requires analytic first and second derivatives of the log- likelihood l defined below. While this approach could also be based on the Poisson assumption defined above for disease rates h_{ij}, a binomial assumption was made for reasons of convenience. We need a little more notation.

Denote $n_{0j} = d_{.j} - n_{.j}$ as the number of incident cases with level j of $S(S = s_j)$ but unknown X;m_{ij} as the number of non-diseased persons with level i of $X(X = x_i)$ and $S = s_j$; $m_{0j} = t_{.j} - d_{.j} - m_{.j}$ as the (typically large) number of non-diseased subjects with $S = s_j$ but unknown i; and q_{ij} as $Pr(X = x_i|S = s_j)$ in the *population*. The complete log-likelihood l is the sum $\sum_{j=1}^{J} l_j$ of stratum-specific log-likelihood l_j, where

$$l_j = \sum_{i=1}^{I} n_{ij}\log(p_{ij}) + \sum_{i=1}^{I} m_{ij}\log(1 - p_{ij}) + \sum_{i=1}^{I} (n_{ij} + m_{ij})\log(q_{ij})$$

$$+ n_{0j}\log\left(\sum_{i=1}^{I} p_{ij}q_{ij}\right) + m_{0j}\log\left(\sum_{i=1}^{I} [1 - p_{ij}]q_{ij}\right).$$

In l_j, the first two terms are the contributions of the cases and controls about p_{ij}, and hence the baseline rate and relative risk parameters; the third term is the contribution of the cases and controls about q_{ij}; and the fourth and fifth terms are the contributions of the diseased and non-diseased subjects with no information about i to the probability of disease. Note that $\sum_{i=1}^{I} p_{ij}q_{ij}$ is the average of the i-specific probabilities of the disease in stratum j, with weights equal to the frequency of level i. Estimates of the variances of the parameter estimates are obtained using the inverse of the observed information at the maximum likelihood estimate.

For this analysis, we estimated $(I - 1)$ separate values of q_{ij} for each j, constrained so that $\sum_{i=i}^{I} q_{ij} = 1$. In our data set there were three values of i,j such that $n_{ij} + \mathrm{m}_{ij} = 0$; we assumed $q_{ij} = 0$ for these i,j. Otherwise, we reparameterized the q's as $\theta_{ij} = \log(q_{ij}/q_{1j})$ for $i = 2, \ldots, I$.

Approach 3 is more difficult to implement than the others because of the complexity of the likelihood and the large number of nuisance parameters ($(I - 1) \times J - 3$ in this example). However, if we restrict our attention to the logistic model, a simplified version of approach 3 can be implemented without the need to estimate nuisance parameters. This is analogous to method (c) in Greenland.[6] A prospective logistic model (equation (1)) fitted to the case-control data yields maximum-likelihood estimates of the relative risk parameters β. Maximum-likelihood estimates of baseline incidence rates are obtained by adding to the stratum parameter estimates $\hat{\alpha}_j$ a term corresponding to the logarithm of the ratio of sampling fractions among controls and cases in stratum j, that is, the quantity $a_j = \log[\{m_{.j}/(t_{.j} - d_{.j})\}/\{n_{.j}/d_{.j}\}]$.[21] They are given by $\hat{h}_{1j} = \{1 + \exp(-\hat{\alpha}_j - a_j)\}^{-1}$. The covariance matrix of estimates of baseline incidence rates and relative risks can then be obtained as described in Section 4 of Prentice and Pyke.[21]

As with approach 2, baseline incidence rates and relative risks are jointly estimated with approach 3, suggesting a potential gain in efficiency relative to approach 1. Further, approach 3 has the advantage of yielding fully efficient maximum likelihood estimates rather than maximum pseudo-likelihood estimates of parameters ξ as approach 2 does.

4. RESULTS

The same form of the model was fit whether maximizing the logistic likelihood in approach 1, the Poisson pseudo-likelihood in approach 2 or the likelihood in approach 3. We estimated 13 parameters for the stratification and exposure variables. We used nine parameters to fit the main effects and interactions of the stratification variables (equation (1)), namely age and geographic area (both in three categories). We used three and one parameters, respectively, to fit the main effects of the exposure variables, smoking (in four categories) and occupational exposure (in two categories). We did not include interaction terms between stratification and exposure variables, or between exposure variables. In particular, since we did not include age × exposure

Table II. Estimates of bladder cancer crude and baseline incidence rates in nine strata

		Estimates of crude incidence rates per 10^5 person years (SDs in parentheses)	Estimates of baseline incidence rates per 10^5 person years (SDs in parentheses)		
			Approach 1	Approach 2	Approach 3
Utah					
age (years)	21–44	3·8 (1·3)	2·2 (0·8)	1·6 (0·6)	1·7 (0·6)
	45–64	32·8 (5·8)	15·1 (3·3)	13·7 (3.0)	13·9 (3·0)
	65+	174·0 (20·7)	91·6 (13.6)	88·4 (14·6)	89·0 (14·0)
Atlanta					
age (years)	21–44	3·6 (1·2)	1·2 (0·4)	1·5 (0·6)	1·4 (0·5)
	45–64	48·1 (6·6)	19·5 (3·2)	18·4 (3·4)	18·8 (3·5)
	65+	224·2 (26·6)	95·8 (13·7)	102·7 (17·0)	101·2 (17·1)
New Jersey					
age (years)	21–44	13·0 (2·1)	5·8 (1·2)	5·4 (1·1)	5·5 (1·1)
	45–64	142·5 (8·1)	55·9 (4·4)	61·5 (7·4)	59·2 (7·3)
	65+	579·8 (25·2)	285·7 (18·7)	276·3 (31·1)	281·9 (31·4)

interaction, we assumed proportional hazards for the effect of smoking and occupational exposure. In approach 3, there were 60 additional nuisance parameters (see Section 3.4).

Table II gives estimates of crude and baseline incidence rates of bladder cancer. They increase sharply with age. As stated before, crude rates are the highest in New Jersey and the lowest in Utah. By multiplying crude rate estimates by the term $1 - \widehat{AR}_j$, we obtained estimates of baseline incidence rates (approach 1). Attributable risks were higher in Atlanta (between 0·57 and 0·67) than in Utah (between 0·42 and 0·54). As a result, baseline incidence rates were nearly equal in these two areas. Estimates of baseline incidence rates with approach 2 were obtained directly by maximizing the Poisson pseudo-likelihood and with approach 3 by maximizing the binomial likelihood. The estimates from the three approaches agree closely. Estimates of standard deviations also agree closely except for two strata (45–64 and 65+ years) in New Jersey with the highest numbers of incident cases, for which standard deviations are markedly smaller with approach 1 than with approaches 2 and 3.

Estimates of relative risks for smoking and occupational exposure are given in Table III. Point estimates and their standard deviations are nearly identical for all three approaches. Note that relative risk estimates from approaches 1 and 3 are maximum likelihood estimates. With approach 2, the observed information matrix $(-\partial E/\partial \xi)^{-1}$ gives variance estimates that would be correct if the terms t_{ij} and d_{ij} were observed rather than estimated, that is if exposure were observed for all individuals in the population. It is interesting to note that the ratios of the correct sandwich variance estimates for the four relative risk parameters, to those obtained from $(-\partial E/\partial \xi)^{-1}$ fall between 0·42 and 0·48, close to the value 0·45 obtained from the usual relative efficiency formula $k/(k+1) \times \pi$, given by Ury,[22] where k is the ratio of the number of controls to cases ($k = 2{\cdot}0$ in the study) and π is the proportion of cases with information on exposure ($\pi = 0.67$ in the study) (see Table I).

Table IV gives estimates of exposure-specific incidence rates for current light smokers (< 20 cigarettes/day) with no occupational exposure and for individuals who are both current heavy smokers ($\geqslant 20$ cigarettes/day) and occupationally exposed. Again, incidence rates estimated by all three approaches agree very closely in all strata. However, there is a gain in efficiency with approaches 2 and 3 since for all strata except stratum 4 (Atlanta, 21–44 years) approaches 2 and

Table III. Estimates of relative risks for smoking and occupational exposure (SDs in parentheses)

	Approach 1	Approach 2	Approach 3
Smoking			
ex-smoker	0·330 (0·125)	0·365 (0·126)	0·330 (0·125)
current light smoker (< 20 cigs/day)	0·633 (0·202)	0·622 (0·207)	0·633 (0·202)
current heavy smoker (⩾ 20 cigs/day)	1·082 (0·127)	1·075 (0·128)	1·082 (0·127)
Occupational exposure	0·449 (0·096)	0·430 (0·099)	0·449 (0·096)

Table IV. Estimates of bladder cancer exposure-specific incidence rates per 10^5 person years for two levels of exposure and in nine strata (SDs in parentheses)

		Current light smokers (< 20 cigs/day) not occupationally exposed			Current heavy smokers (⩾ 20 cigs/day) occupationally exposed		
		Approach 1	Approach 2	Approach 3	Approach 1	Approach 2	Approach 3
Utah age (years)	21–44	4·1 (1·7)	2·9 (1·2)	3·2 (1·3)	10·0 (3·9)	7·1 (2·5)	8·0 (2·8)
	45–64	28·4 (8·2)	25·6 (6.9)	26·1 (6·9)	69·8 (17·5)	61·9 (12·0)	64·1 (12·4)
	65+	172·6 (43·0)	164·7 (40·6)	167·4 (38·1)	423·5 (84·5)	398·5 (62·1)	409·9 (61·7)
Atlanta age (years)	21–44	2·2 (0·9)	2·9 (1·3)	2·7 (1·1)	5·4 (2·0)	6·9 (2·7)	6·6 (2·4)
	45–64	36·7 (9·4)	34·3 (8·6)	35·5 (8·6)	90·0 (19·4)	83·0 (13·1)	86·9 (13·6)
	65+	180·3 (44·9)	191·2 (43·1)	190·5 (41·7)	442·6 (90·0)	462·6 (72·2)	466·1 (69·2)
New Jersey age (years)	21–44	11·0 (3·1)	10·1 (2·7)	10·4 (2·6)	27·0 (6·5)	24·5 (4·5)	25·4 (4·6)
	45–64	105·3 (22·6)	114·4 (22·1)	111·5 (21·0)	258·4 (42·0)	276·9 (26·0)	273·1 (23·9)
	65+	538·0 (111·3)	514·4 (100·0)	530·0 (99·3)	1320·3 (204·7)	1245·0 (116·0)	1289·8 (117·9)

3 yield standard deviation estimates smaller than those obtained from approach 1. Relative to approach 1, the standard deviation estimates are between 2 per cent and 29 per cent (respectively, 7 per cent and 24 per cent) smaller with approach 2 (respectively, approach 3) for current light smokers in all strata but stratum 4 and between 20 per cent and 43 per cent (respectively, 22 per cent and 43 per cent) smaller in individuals who are current heavy smokers and occupationally exposed in the same eight strata. This is similarly observed for all levels of exposure other than the baseline and is due to the negative correlations between baseline incidence rates and relative risk estimates. With approach 1, these correlations only come from part of baseline rate estimates, namely the term $1 - \widehat{AR}_j$, while with approaches 2 and 3, baseline rates and relative risks are estimated jointly and are thus more strongly correlated. On these data, approaches 2 and 3 are about equally efficient.

5. DISCUSSION

The ability to make inference about exposure-specific incidence rates is an advantage of population-based case-control studies[23] that has been long recognized,[1,2] but not fully exploited in the multivariate setting until recently.[7] Accurate estimates of exposure-specific incidence rates depend on having a complete and accurate census of the cases. The three approaches to analysis presented here offer a choice of methods that can be used in a multivariate setting. Approaches 2 and 3 are closely related to work by Scott and Wild[24] and Wild[25] on fitting prospective regression models for case-control studies. Those papers, however, concentrated on estimation of relative risk.

All three approaches require the assumption that data be missing at random. Violation of this assumption will have consequences for our approaches. They also affect estimation of relative risk parameters in standard case-control studies. The fact that interviewing response rates vary by disease status, age and geographic area, does not imply the missing at random assumption is violated.

The bladder cancer exposure-incidence rates that we obtained with the three approaches agree very closely. Approaches 2 and 3 appear to be more efficient. These approaches rely more heavily on the relative risk model than approach 1. Therefore, these approaches may be more sensitive to model misspecification than approach 1, but less variable than approach 1 when the model is correct. Further numerical work and simulations are needed to investigate this supposition. Simulations are also needed to confirm the findings on efficiency in this paper. In particular, it is necessary to confirm that our variance estimators accurately reflect sample variance.

An advantage of approaches 2 and 3 is that they can be applied to more general models of risk. For instance, it would be possible to assume an additive form:

$$h_{ij} = h_{1j} + D_{ij},$$

where D_{ij} is the difference between incidence rates at levels i and 1 of X in stratum j. An advantage of approach 1 is that it can be applied to data with individual matching of the controls to the cases, as well as with controls selected by frequency matching as assumed in this paper. Another advantage of approach 1 is that if external crude rates are available, it can be applied to case-control studies that are not population-based.

From estimates of exposure-specific incidence rates, one can estimate the probability that an individual at a given level of exposure will develop disease over a specified age interval. Such probabilities have been called 'absolute risks'[7] and have been used to assist in medical counselling[7] and to plan prevention trials to assess interventions.[26]

One purpose of the National Bladder Cancer Study was to see whether risk factors might explain extreme geographic differences in age-adjusted bladder cancer rates. Because controls were frequency matched on geographic area, routine methods for estimating the effect of area do not apply. Our methods, however, can be used. In the oldest age group, the crude incidence rates per 100,000 person years are 174,224 and 578 (Table II) for white males in Utah, Atlanta and New Jersey, respectively. After adjustment, the baseline rates are estimated as 92, 96 and 286 (Table II), suggesting that the difference between Utah and Atlanta might be explained by the two risk factors discussed here, namely smoking and occupational exposure, but that the rate in New Jersey is still far higher than in the other two areas.

The analyses presented here are preliminary and are presented only as an example of application of the methods we have developed. We ignored many exposure factors, did not adjust finely for age, smoking, or occupational exposure, and only considered white males in three of ten available areas. We are planning a more complete analysis of the National Bladder Cancer Study in order to obtain more extensive exposure-specific incidence rates based on all exposure factors described in Section 2 for all 10 geographic areas and all race-sex groups.

REFERENCES

1. Cornfield, J. 'A method for estimating comparative rates from clinical data: Applications to cancer of the lung, breast and cervix', *Journal of the National Cancer Institute*, **11**, 1269–1275 (1951).
2. Cornfield, J. 'A statistical problem arising from retrospective studies', *in* Neyman, J. (ed.), *Proceedings of the Third Berkeley Symposium, Volume IV*, University of California Press, Monterey, 1956, pp. 133–148.

3. MacMahon, B. 'Prenatal X-ray exposure and childhood cancer', *Journal of the National Cancer Institute*, **28**, 1173–1191 (1962).
4. Miettinen, O. S. 'Proportion of disease caused or prevented by a given exposure, trait or intervention', *American Journal of Epidemiology*, **99**, 325–332 (1974).
5. Neutra, R. R. and Drolette, M. E. 'Estimating exposure-specific disease rates from case-control studies using Bayes' theorem', *American Journal of Epidemiology*, **108**, 214–222 (1978).
6. Greenland, S. 'Multivariate estimation of exposure-specific incidence from case-control studies', *Journal of Chronic Diseases*, **34**, 445–453 (1981).
7. Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C. and Mulvihill, J. J. 'Projecting individualized probabilities of developing breast cancer for white females who are being examined annually', *Journal of the National Cancer Institute*, **81**, 1879–1886 (1989).
8. Hartge, P., Cahill, J. J., West, D., Hauck, M., Austin, D., Silverman, D. and Hoover, R. 'Design and methods in a multi-center case-control interview study', *American Journal of Public Health*, **74**, 52–56 (1985).
9. Silverman, D. T., Levin, L. I., Hoover, R. N. and Hartge, P. 'Occupational risks of bladder cancer in the United States. I. White men', *Journal of the National Cancer Institute*, **81**, 1472–1480 (1989).
10. Miller, B. A., Ries, L. A. G., Hankey, B. F., Kosary, C. L. and Edwards, B. K. *Cancer Statistics Review: 1973–1989*, National Cancer Institute, NIH publication no. 92-2789, 1992.
11. Breslow, N. E. and Day, N. E. *Statistical Methods in Cancer Research. Vol. II. The Design and Analysis of Cohort Studies*, IARC Scientific Publications, Lyon, 1987.
12. Holford, T. R. 'The analysis of rates and of survivorship using log-linear models', *Biometrics*, **36**, 299–305 (1980).
13. Laird, N. and Oliver, D. 'Covariance analysis of censored survival data using log-linear analysis techniques', *Journal of the American Statistical Association*, **76**, 231–240 (1981).
14. Little, R. J. A. and Rubin, D. B. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
15. Bruzzi, P., Green, S. B., Byar, D. P., Brinton, L. A. and Schairer, C. 'Estimating the population attributable risk for multiple risk factors using case-control data', *American Journal of Epidemiology*, **122**, 904–914 (1985).
16. Benichou, J. and Gail, M. H. 'Methods of inference for estimates of absolute risk derived from population-based case-control studies', *Biometrics* (to appear).
17. Rao, C. R. *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.
18. Benichou, J. and Gail, M. H. 'A delta-method for implicitly defined random variables', *The American Statistician*, **43**, 41–44 (1989).
19. Liang, K. Y. and Zeger, S. L. 'Longitudinal data analysis using generalized liner models', *Biometrika*, **73**, 13–22 (1986).
20. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood estimates from incomplete data via the EM algorithm', *Journal of The Royal Statistical Society, Series B*, **39**, 1–38 (1977).
21. Prentice, R. L. and Pyke, R. 'Logistic disease incidence models and case-control studies', *Biometrika*, **66**, 403–411 (1979).
22. Ury, H. 'Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data', *Biometrics*, **31**, 643–649 (1975).
23. Wacholder, S., Silverman, D. T., McLaughlin, J. K. and Mandel, J. S. 'Selection of controls in case-control studies. II. Types of controls', *American Journal of Epidemiology*, **9**, 1029–1041 (1992).
24. Scott, A. J. and Wild, C. J. 'Fitting logistic regression models in stratified case-control studies', *Biometrics*, **47**, 497–510 (1991).
25. Wild, C. J. 'Fitting prospective regression models to case-control data', *Biometrika*, **78**, 705–717 (1991).
26. Smigel, K. 'Breast cancer prevention trial takes off', *Journal of the National Cancer Institute*, **84**, 669–670 (1992).